Sender Continuation Perplexity:
Measuring Reasoning Trajectory Alignment at the Natural Language Boundary

Jeremy McEntire[1]

March 2026

Abstract

Papers XVI–XVII established that domain classification saturates at 95% from text alone, and that activation injection preserves geometric structure (RSA = 0.47) that text destroys (RSA = 0.11). Neither metric directly measures whether a receiving agent can follow the sender's specific reasoning chain. This paper introduces sender continuation perplexity: feed agent 1's generated tokens to agent 2 (with various coordination signals) and measure per-token cross-entropy. Low perplexity means agent 2 expects agent 1's specific word choices — a direct probe of reasoning trajectory alignment.

Agent 2 with the original context achieves PPL = 1.59 (near-deterministic prediction). Text summary achieves PPL = 3.05, preserving 63.7% of the perplexity gap. INLP injection adds 0.3% and full activation injection adds 1.2% over text. Critically, **full activation injection without text scaffold (BOS + full) achieves PPL = 5.60, indistinguishable from no context at all** (PPL = 5.62). Despite preserving 4× higher geometric fidelity (Paper XVII), activation injection without text carries no reasoning trajectory information.

Text is the primary carrier of reasoning trajectory alignment. Activation injection provides geometric structure that the forward pass cannot translate into functional state. The 36.3% perplexity gap between text summary and original context represents the information ceiling for text-based coordination — and activation injection at a single layer does not close it.

1 Introduction

Papers XVI–XVII used domain classification and RSA to measure coordination fidelity. Both metrics probe representational structure — the geometry of activations at the terminal layer. Neither directly measures the quantity most relevant to multi-agent coordination: whether a receiving agent can continue the sender's line of reasoning.

Continuation perplexity provides this measurement. Agent 1 processes a domain probe and generates a 64-token continuation. Agent 2, equipped with various coordination signals

[1] Correspondence: jmc@cageandmirror.com

(text summary, INLP injection, full activation injection), is asked to predict agent 1's tokens. Per-token cross-entropy measures how "surprised" agent 2 is by agent 1's specific choices. Low perplexity indicates that agent 2 has recovered agent 1's reasoning trajectory, not just its domain identity.

The KV-cache implementation is critical: agent 2's context is processed first (with optional injection at layer 10), producing key-value pairs that encode the coordination signal. Agent 1's continuation tokens are then teacher-forced through the cached attention mechanism. This ensures that the injection at the last context token propagates through the attention to affect all continuation token predictions.

2 Methods

2.1 Sender continuation generation

Each of 160 domain probes (40 per domain) is processed by Qwen 2.5-7B. The model generates a greedy 64-token continuation. Layer-10 activations are captured during a separate forward pass of the probe text. INLP directions (36, 9 per domain) are computed at layer 10 via iterative ridge regression.

2.2 Conditions

Seven conditions probe different coordination mechanisms:

1. **Original context**: Agent 2 sees the exact same probe text as agent 1. Upper bound.

2. **Text summary**: Agent 2 sees a 3-sentence summary of the probe. Text-only baseline.

3. **Text + INLP (36d)**: Summary + INLP projection of sender's centered layer-10 activation, injected at layer 10 ($\alpha = 1.0$).

4. **Text + Full (3584d)**: Summary + full centered layer-10 activation, injected at layer 10.

5. **BOS + Full (3584d)**: Single BOS token + full activation injection. No text scaffold.

6. **No context**: BOS token only. Lower bound (excluding noise floor).

7. **Scrambled**: Text summary context, but agent 1's continuation tokens are randomly shuffled. Noise floor.

2.3 Perplexity computation

For each probe and condition:

1. Process the context (with optional injection) to build a KV cache.
2. Teacher-force agent 1's 64 continuation tokens using the cached key-value pairs.
3. Compute per-token cross-entropy: $\mathrm{CE}_t = -\log P(x_t \mid x_{<t}, \text{context})$.
4. Report mean CE and perplexity = exp(mean CE).

The KV-cache approach ensures that injection at layer 10 of the context forward pass propagates through the attention mechanism to affect continuation token predictions.

3 Results

3.1 Primary result: text preserves 63.7% of reasoning trajectory

Table 1: Continuation perplexity across coordination conditions.

Condition	PPL	CE	Δ vs Text	Improvement
Original context	**1.59**	0.445	—	Upper bound
Text summary	3.05	1.030	—	63.7% of gap
Text + INLP (36d)	3.05	1.029	−0.001	+0.3% over text
Text + Full (3584d)	3.03	1.025	−0.005	+1.2% over text
BOS + Full (3584d)	5.60	1.482	+0.452	≈ No context
No context	5.62	1.483	+0.453	Lower bound
Scrambled	1637	7.111	—	Noise floor

The original context (PPL = 1.59) is nearly deterministic: agent 2 predicts agent 1's tokens with high confidence when given the same input. Text summary (PPL = 3.05) preserves 63.7% of the perplexity gap between original and no-context conditions. The remaining 36.3% represents information that text compression loses.

INLP injection adds 0.3% improvement over text (CE: 1.030 → 1.029). Full activation injection adds 1.2% (CE: 1.030 → 1.025). Both improvements are real (the KV-cache ensures the injection propagates) but negligible relative to the text contribution.

3.2 BOS + Full activation: geometric preservation without functional alignment

BOS + Full achieves PPL = 5.60, indistinguishable from no context (PPL = 5.62). This is the most striking result: Paper XVII showed that BOS + Full achieves RSA = 0.47 (4× higher than text). Despite preserving the sender's representational geometry with high fidelity, this geometric information carries *zero* reasoning trajectory alignment.

The forward pass transforms the injected activation through 17 layers of attention (layers 10–27) and the output head. This transformation preserves the pairwise geometric structure (which probes are similar to which) but does not preserve the specific processing state that determines next-token prediction. RSA and perplexity measure fundamentally different quantities.

3.3 Per-domain breakdown

Table 2: Per-domain perplexity for text summary and original context conditions.

Condition	Medical	Legal	Code	Science
Original context	—	—	—	—
Text summary	2.60	3.04	3.05	2.49
BOS + Full	3.96	4.94	4.93	3.88
No context	3.97	4.95	4.93	3.88

Medical and science achieve lower text-baseline perplexity (2.5–2.6) than legal and code (3.0–3.1), suggesting that medical and science continuations are more predictable from summaries. This is notable given Paper XVII's finding that science has the *lowest* within-domain RSA for text (near zero). Text preserves science's functional trajectory (low perplexity) despite destroying its geometric structure (low RSA). The two types of preservation are independent.

The BOS + Full condition shows identical per-domain perplexity to no-context, confirming that the activation injection provides no domain-specific reasoning information when text scaffold is absent.

4 Discussion

4.1 The perplexity gap as information ceiling

The 36.3% perplexity gap between text summary (PPL 3.05) and original context (PPL 1.59) represents the information ceiling for text-based coordination. This gap contains:

- **Epistemic state**: Agent 1's specific beliefs about the problem, not just the domain's general knowledge.
- **Reasoning trajectory**: Which inference path agent 1 took to reach its current state.
- **Uncertainty topology**: Which parts of the problem agent 1 considers resolved vs. open.
- **Conceptual specificity**: Not "medical analysis" but "this specific diagnostic pathway at this point in the chain."

Single-layer activation injection (INLP or full) closes $\leq 1.2\%$ of this gap. The injected activation at layer 10 encodes a geometric snapshot of the sender's processing state, but the receiver's forward pass cannot translate this snapshot into the specific next-token predictions that would demonstrate reasoning alignment.

4.2 Why geometric preservation fails functionally

Paper XVII showed that BOS + Full preserves representational geometry (RSA = 0.47) while text destroys it (RSA = 0.11). Yet text achieves far better perplexity (3.05 vs 5.60). The explanation lies in the asymmetry between RSA and perplexity:

- RSA measures *relative* structure: which probes are similar to which. It is invariant to the absolute position in activation space.
- Perplexity measures *absolute* prediction: does the model predict *this specific token*? It requires the model's internal state to converge on the exact same processing trajectory.

Activation injection at layer 10 preserves relative structure (the geometry of the domain space) but does not place the receiver in the same *absolute* processing state as the sender. Text does the opposite: it places the receiver in a similar functional state (low perplexity) through completely different geometric means (low RSA). Functional alignment is achieved through semantic content, not geometric alignment.

4.3 Implications for activation-sharing protocols

1. **Text is the primary reasoning channel.** For multi-agent coordination where agents need to follow each other's reasoning, text summaries are far more effective than activation injection.

2. **Single-layer injection is insufficient.** Injecting at one layer provides geometric structure that the forward pass dilutes over 17 subsequent layers. Multi-layer injection or terminal-layer injection may be necessary to achieve functional alignment.

3. **Activation-level communication has a role, but not for reasoning.** The geometric structure preserved by activation injection (Paper XVII) may be useful for tasks where relative similarity matters (routing, clustering, ensemble weighting) rather than tasks requiring specific reasoning chains.

4. **The 36.3% gap is the design target.** Any coordination protocol that improves on text-only must close this gap. The gap contains the reasoning trajectory information that text compression loses and single-layer injection cannot recover.

5 Conclusion

Sender continuation perplexity directly measures reasoning trajectory alignment between cooperating model instances. Text summaries preserve 63.7% of this alignment. Activation injection at layer 10 adds at most 1.2%. Full activation injection without text scaffold preserves geometric structure (RSA = 0.47, Paper XVII) but carries zero reasoning trajectory information (PPL = 5.60 ≈ no context).

The dissociation between geometric preservation and functional alignment is now established across three measurement levels: domain classification (Paper XVI), representational similarity (Paper XVII), and continuation perplexity (this paper). Text preserves function without geometry; activations preserve geometry without function. The 36.3% perplexity gap — containing epistemic state, reasoning trajectory, uncertainty topology, and conceptual specificity — remains the binding constraint on multi-agent coordination.

Data Availability

All results are archived at `huggingface.co/datasets/jmcentire/paper8-data` under `paper18/`.

Series: Activation Geometry of Domain-Selective Noise Injection, Paper XVIII.